¿Podrían los ordenadores detectar emociones cuando les hablamos?

C. Rincon, R. Barra, J. M. Lucas, J. M. Montero, J. Macias-Guarasa, L.F. D'Haro, F. Fernández, R. San-Segundo, J.Ferreiros, R. Córdoba, J.M. Pardo
{carmenr, barra, juanmak, juancho, macias, lfdharo, efhes, lapiz, jfl, cordoba, pardo}@die.upm.es

***Abstract* — Este trabajo describe un conjunto de experimentos sobre identificación de emociones en busca de completar la inteligencia emocional de un robot guía dentro del entorno de un museo. Los experimentos se centran en el reconocimiento de emociones en dos idiomas (castellano y alemán), así como posibilidad de reconocer las emociones independientemente del idioma. Los resultados inducen a que sería factible detectar emociones tales como la tristeza o el enfado del visitante.**

I. INTRODUCCIÓN

En el proyecto ROBINT, se ha diseñado un robot autónomo con capacidad para servir de guía en un museo como el de las Ciencias Príncipe Felipe de Valencia. Dicho robot, además de las capacidades de navegación que le permiten moverse por el recinto, incorpora un reconocedor de habla independiente del locutor, un sistema de comprensión automática de habla y un conversor texto-habla para poder guiar a sus visitantes humanos de una manera lo más natural posible.

Para poder realizar un sistema que resulte aún más empático y atractivo para los visitantes, se ha dotado al robot de capacidad para modular su voz de acuerdo con su estado emocional, siendo así el robot capaz de simular por medio del habla que se encuentra triste, alegre, enfadado o sorprendido, y no sólo en un estado neutro o estándar.

Sin embargo, para poder incorporar las emociones de una manera más integral en el comportamiento del robot, sería necesario dotarlo de capacidad para detectar emociones en la voz de sus visitantes y, sobre todo, en la voz del personal del museo que lo acompañe o supervise. Por ello en este artículo hemos llevado a cabo experimentos de identificación de emociones en la voz.

II. DESCRIPCIÓN DE LAS BASES DE DATOS

Para la evaluación del sistema desarrollado emplearemos dos bases de datos actuadas en diferentes idiomas, castellano y alemán. La razón por la que empleamos estas bases de datos es para intentar descubrir características propias de la emoción manifestadas en la voz, que sean independientes del idioma.

A. Spanish Emotional Speech (SES)

SES es la base de datos en castellano, descrita en [2]. Está formada por tres sesiones de grabación de habla con emociones interpretadas por un único actor masculino. Cada sesión de grabación incluye treinta palabras (2 minutos), quince frases cortas (7 minutos) y cuatro párrafos (39 minutos), interpretando cuatro emociones primarias (alegría, enfado en frío y sorpresa) y voz interpretada según el estado neutro. El texto interpretado no posee ningún contenido emocional intrínseco. Esta base de datos fue parcialmente etiquetada fonética y prosódicamente de forma semiautomática.

B. Berlin Database of Emotional Speech (EMODB)

EMODB es la base de datos en alemán, descrita en [3]. Está formada por una o dos sesiones de grabación de habla con emociones interpretadas por diez actores (cinco hombres y cinco mujeres). Cada sesión incluye diez frases (cinco cortas y cinco largas, obteniendo un total de 24,5 minutos), interpretando seis emociones (alegría, enfado en caliente, tristeza, aburrimiento, asco y miedo) y voz interpretada según el estado neutro. EMODB dispone de 24,5 minutos de voz distribuidos de manera no homogénea entre los distintos actores y las distintas emociones. Esta base de datos está etiquetada fonéticamente, pero no prosódicamente.

III. DESCRIPCIÓN DEL SISTEMA

El sistema automático de identificación de emociones empleado, lo podemos dividir en cuatro fases principales. La primera de ellas sería la de *parametrización*, en la cual se extraeran características representativas de la voz. El estado emocional del individuo es transmitido parcialmente mediante cambios en la expresión facial. Estos cambios traen consigo la modificación del tracto vocal, provocando así variaciones en la señal de voz. Debido a esto, hemos utilizado los MFCC (Mel Frecuency Cepstrum Coefficients), ampliamente utilizados en tecnología del habla, como características representativas del tracto vocal.

La siguiente fase es opcional, se trata de la *normalización* de dichas características, intentando reducir la variabilidad inter-locutor y del canal de audio. Esta normalización resulta interesante en los vectores obtenidos a partir de la base de datos EMODB, que es multilocutor. Aplicaremos técnicas de normalización basadas en la media (CMN – Cepstral Mean Normalization), en la varianza (CVN – Cepstral Variance Normalization) o en ambas (CMN + CVN). En la fase de *entrenamiento*, se genera un modelo para cada una de las emociones a partir de los vectores de características. El entrenamiento está basado en un modelo de mezcla de gausianas (GMM – Gaussian Mixture Model) [4]. Finalmente, en la fase de *clasificación*, se decide a qué emoción pertenece un ejemplo de voz recibido como entrada del sistema. Se ha

utilizado un clasificador bayesiano en el que se estima la probabilidad (acumulada sobre todas las tramas que componen dicho ejemplo de voz) de que pertenezca a las distintas emociones que forman parte del sistema.

IV. EXPERIMENTACIÓN

Se han realizado experimentos de identificación de emociones en castellano y en alemán por separado. Adicionalmente, se ha estudiado la semejanza entre emociones de ambos idiomas; utilizando los modelos de emociones en castellano y tratando de identificar las emociones en alemán.

A. Identificación de las emociones en Castellano

En la interacción con un usuario el robot podría identificar el estado emocional pasadas varias interaciones entre el robot y la persona; o bien tratar de identificar dicho estado en cada interacción. Dado que en SES disponemos de distintos tipos de locuciones, párrafos y frases cortas, podemos aproximarnos a los dos escenarios según tratemos de identificar la emoción basándonos en párrafos o en frases.

Los resultados obtenidos cuando los modelos de emociones son entrenados con frases cortas y se identican las emociones sobre los párrafos es de un 97,62%. Este resultado nos indica la viabilidad de reconocer el estado emocional, conocido el individuo, una vez realizadas varias interacciones con él.

Si por el contrario, se trata de identificar la emoción basándonos en frases cortas (los modelos son entrenados con los párrafos), la tasa media de identificación es 88,86%. La matriz de confusión que obtenemos cuando empleamos la frase como unidad de clasificación, es la que se muestra en la siguiente tabla:

TABLA I
MATRIZ DE CONFUSIÓN PARA LAS EMOCIONES DE SES (VALORES EN %)

Emoción interpretada	**Emoción Identificada**				
	Alegría	**Enfado en frío**	**Sorpresa**	**Tristeza**	**Neutra**
Alegría	**80,53**	2,31	15,51	0,99	0,66
Enfado en frío	2,31	**92,08**	2,31		3,3
Sorpresa	10,23	0,33	**89,44**		
Tristeza		0,66		**93,4**	5,94
Neutra	3,69			4,15	**92,17**
PRECISIÓN	*83,23*	*96,54*	*83,38*	*94,79*	*90,3*

Las emociones que mejor se identifican son la tristeza, el enfado y la neutra. Estas emociones obtienen tanto tasa de identificación como precisión elevadas. Por otro lado, las tasas de identificación de la sorpresa (89,44%) y principalmente la alegría (80,53%) son las más bajas, debido mayoritariamente a la confusión que existe entre ambas emociones, de naturaleza positiva. Este hecho, observado en experimentos perceptuales previos [2], implica un descenso de la precisión de estas dos emociones. Así mismo se produce cierta confusión entre la tristeza y la neutra, las dos emociones con menor nivel de activación.

B. Identificación de las emocionas en alemán

En este caso, nos encontramos con una tarea multilocutor y la identificación de 7 emociones. En este experimento se ha entrenado el sistema con nueve de los diez locutores y se han clasificado las emociones del locutor restante. La tasa de identificación final (46,06%) se ha calculado como el promedio de la tasa para cada uno de los locutores.

Observamos que se trata de una tasa mucho menor que la obtenida al emplear datos de SES. No obstante, los resultados perceptuales para esta base de datos (86%) son menores que los obtenidos para SES (90%), se dispone de menos datos y sin embargo más emociones y mayor variabilidad, al ser una base de datos multilocutor. Con el objetivo de aumentar la tasa de identificación, aplicaremos las distintas normalizaciones comentadas en el apartado previo. Los resultados y mejoras relativas de aplicar dichas técnicas se muestran en la siguiente tabla:

TABLA II
TASA MEDIA DE IDENTIFICACIÓN Y GANANCIA RELATIVA EN FUNCIÓN DE SI NORMALIZAMOS O NO LOS VECTORES DE CARACTERÍSTICAS DE EMODB (VALORES EN %)

	Sin Normalizar	**CMN**	**CVN**	**CMN + CVN**
Tasa media de identificación	46,06	57,46	51,48	59,18
Ganancia relativa		21,1	10	24,3

Observamos en la tabla que la normalización con la que obtenemos mejores resultados es CMN + CVN (24,3% de mejora relativa). A continuación se presentan de forma detallada los resultados obtenidos para cada emoción aplicando este tipo de normalización .

TABLA III
MATRIZ DE CONFUSIÓN PARA LAS EMOCIONES DE EMODB CON CMN + CVN (VALORES EN %)

Emoción interpretada	Emoción Identificada						
	Alegría	**Enfado en caliente**	**Aburrimiento**	**Tristeza**	**Asco**	**Miedo**	**Neutra**
Alegría	**42,25**	26,76			5,63	23,94	1,41
Enfado en caliente	13,39	**71,65**			0,79	14,17	
Aburrimiento			**43,21**	12,35	12,35	6,17	25,93
Tristeza			1,61	**95,16**			3,23
Asco	2,17		2,17		**73,91**	13,04	8,7
Miedo	8,7	13,04	4,35	7,25	11,59	**44,93**	10,14
Neutra			36,71	3,8	11,39	5,06	**43,04**
PRECISIÓN	*63,52*	*64,29*	*49,07*	*80,26*	*46,55*	*63,9*	*41,87*

La emoción que mejor se identifica es la tristeza (95,16%), seguida del asco (73,91%) y el enfado (71,65%). Sin embargo, la precisión del asco es excesivamente baja (46,55%); todas las emociones, salvo la tristeza se confunden con ella. Si analizamos en detalle la matriz de confusión, podemos dividir las emociones en dos grandes grupos, atendiendo a la similitud que existe entre ellas, según su grado de confusión. Así, la alegría, el enfado y el miedo, emociones con un nivel de activación elevado, se confunden principalmente entre ellas (en el caso de miedo también se confunde con el asco). Por el contrario, emociones "apagadas" como la neutra, la tristeza o el aburrimiento se confunden entre sí; encontrándose el asco en medio de estas dos agrupaciones.

C. Identificación de emociones entre distintos idiomas

A fin de intentar descubrir posibles semejanzas entre las emociones en castellano y en alemán, dado que estamos tratando de emociones conocidas como primarias [1], se han clasificado las emociones en alemán utilizando los modelos entrenados en castellano. En este caso sólo podremos emplear las emociones que son comunes a ambas bases de datos, que son la alegría, el enfado (aunque como veremos, no es despreciable la naturaleza intrínseca del mismo), la tristeza y la neutra. En la siguiente tabla se muestran las tasas de identificación obtenidas y las mejoras relativas obtenidas al emplear las distintas normalizaciones:

TABLA IV
TASA MEDIA DE IDENTIFICACIÓN Y GANANCIA RELATIVA EN FUNCIÓN DE SI NORMALIZAMOS O NO LOS VECTORES DE CARACTERÍSTICAS DE SES, CLASIFICANDO EMODB (VALORES EN %)

	Sin Normalizar	**CMN**	**CVN**	**CMN + CVN**
Tasa media de identificación	46,61	48,79	48,62	51,56
Ganancia relativa		5,1	4,8	11,7

La tasa de identificación obtenida empleando vectores de características sin normalizar, es ligeramente mayor de la que obteníamos cuando empleábamos solamente datos en alemán (46,61% frente a 46,06%), aunque esta diferencia no es estadísticamente significativa. Sin embargo, en este caso la mejora obtenida al emplear vectores normalizados es menor. Los mejores resultados se obtienen cuando normalizamos respecto a la media y la varianza (11,7% frente al 24,3% del caso anterior). Las tasas de identificación para cada emoción empleando esta normalización son las que se muestran en la siguiente tabla:

TABLA V
MATRIZ DE CONFUSIÓN DEL EXPERIMENTO EN EL QUE ENTRENAMOS CON SES Y CLASIFICAMOS EMODB, EMPLEANDO LA NORMALIZACIÓN CMN + CVN (VALORES EN %)

Emoción interpretada	Emoción Identificada			
	Alegría	**Enfado en frío**	**Tristeza**	**Neutra**
Alegría	**77,46**	21,13		1,41
Enfado en caliente	77,17	**22,83**		
Tristeza			**100**	
Neutra	12,66	10,13	41,77	**35,44**
PRECISIÓN	*46,31*	*42,22*	*70,54*	*96,18*

En este caso conseguimos identificar siempre la tristeza, pero la neutra se confunde con ella en un 41,77%, haciendo que disminuya su precisión. La tasa de identificación obtenida para la alegría es elevada (77,46%), pero el enfado se confunde en un 77,17% con ella, lo que hace que la precisión obtenida para la alegría sea muy baja (12,66%). Esta confusión producida entre el enfado en alemán y la alegría en castellano se debe a que el enfado interpretado en alemán es un enfado en caliente,

con un elevado grado de activación, similar al de la alegría; mientras que el enfado interpretado en castellano es un enfado en frío con un bajo nivel de activación. Si analizamos en detalle los resultados, se observa como de nuevo existe cierta agrupación entre las distintas emociones, teniendo por una parte la alegría y el enfado; y por la otra, la tristeza y la neutra.

VI. Conclusión

En este trabajo se ha estudiado la posibilidad de identificar el estado emocional de un usuario que interaccione con un robot. Se han estudiado conjuntos de emociones en dos idiomas distintos, así como la posibilidad de identificar dichas emociones independientemente del idioma.

Los resultados obtenidos en castellano (97% cuando se identifica la emoción basándose en varias frases y 92% cuando se identifica la emoción en cada frase) muestran la viabilidad de identificar el estado emocional del individuo conocida la identidad del mismo.

La identificación del estado emocional en una tarea multi-locutor necesitaría de mayor número de datos de entrenamiento. No obstante, los datos que se obtienen, muestran una detección clara de emociones negativas como el enfado (71,6%) o la tristeza (95,2%).

Se ha realizado un estudio inicial sobre la posibilidad de identificar el estado emocional independientemente del idioma del usuario. En este caso, la tristeza es claramente identificable y emociones como la alegría presentan tasas de identificación elevadas (77,5%).

De los resultados obtenidos se concluye con la existencia de cierta relación entre la voz, concretamente de los MFCC, y características relacionadas con el nivel de activación emocional. Los resultados de identificación muestran dos grupos, en función del nivel de activación de las emociones. Así, emociones activas como la alegría, el enfado en caliente y el miedo se identifican entre ellas; al igual que sucede con emociones de baja actividad como la neutra, la tristeza y el aburrimiento.

Finalmente, los resultados muestran que la tristeza es claramente identificable tanto en castellano como en alemán y la posibilidad de identificarla independientemente del idioma.

AGRADECIMIENTOS

Agradecimientos a Ramón Galán y todo el Grupo de Control Inteligente de la ETSII de la UPM como "padres" del robot al que los autores enseñamos a hablar y escuchar.

Este trabajo ha sido apoyado parcialmente por EDECAN (TIN2005-08660-C04-04), ROBINT (DPI2004-07908-C02-02), ATINA (UPM-CAM REF: CCG06-UPM/COM-516) y TINA (UPM-CAM REF: R05/10922).

REFERENCIAS

[1] R. Cowie and R.R. Cornelious, “Describing the emotional status that are expressed in speech” *Speech Comumunication*, vol.1, no. 40, 2003.

[2] J. M. Montero, J. Gutierrez-Ariola, S. Palazuelos, E. Enriquez, and J. M. Pardo, “Spanish emotional speech from database to tts,” in *Proceedings of ICSLP,* pp. 923-925, September 1998.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Proceedings of Interspeech,* Lissabon, Portugal, September 2005.

[4] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing: A guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 1st edition, April 2001.

[5] R. Barra, J. M. Montero, J. Macías-Guarasa, L- F. Dharo, R. San-Segundo, and R. Córdoba, “Prosodic and segmental rubrics in emotion identification,” in *Proccedings of ICASSP,* September 2006.

[6] Ben Milner and Xu Shao, *Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model”,* School of Information Systems, University of East Anglia, Norwich, UK. 2002.

[7] Richard O. Duda, Meter E. Hart, and David G. Stork, *Pattern Classification*, Second Edition, Wiley-Interscience, 2001.

[8] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, “Reconocimiento automático de emociones utilizando parámetros prosódicos,” in *Procesamiento del Lenguaje Natural,* June 2005.